Data Curation 101

ABCs of Data Curation & Scholarly Communication
Science & Engineering Academic Librarians (SEAL-S) Fall Program
15 October 2010

Jeanine Scaramozzino
College of Science & Mathematics Librarian
School of Education Librarian
Cal Poly State University

VALUE THE SCHOLAR IN EVERYONE

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Overview

- Managing data & metadata
- Stakeholders
- Ethics, legality & copyright
- Data management plans
- Library roles
- Challenges

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Data Curation Defined

"Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarly and educational activities."

GSLIS, UIUC: http://cirss.lis.illinois.edu/CollMeta/dcep.html

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Why manage data?

- Essential raw material of science
- Demand for transparency of science
- Funder requirements
- Publisher recommendations

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Data Management

- Activities
  - Enable discovery & retrieval
  - Maintain data quality
  - Preserve & archive for re-use
- Tasks
  - Appraisal & selection
  - Data integrity
  - Interoperability

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

What is Data?

- Text (e.g. flat text files, Word, PDF)
- Numerical (e.g. SPSS, STATA, Excel, Acess, MySQL)
- Multimedia (e.g. jpeg, tiff, dicom, mpeg, quicktime)
- Models (e.g. 3D, statistical)
- Software (e.g. Java, C)
- Domain-specific (e.g. FITS in astronomy, CIF in chemistry)
- Instrument-specific (e.g. Olympus Confocal Microscope Data Format)

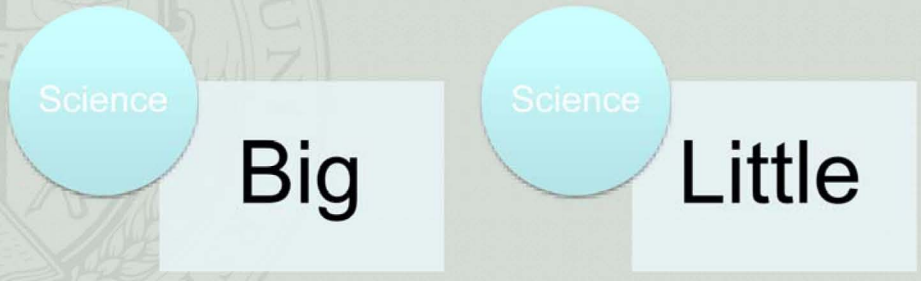
Size of Data Sets

Relevant References

•"Big Science" was originally coined by Alvin M. Weinberg in his 1961 work Impact of Large-Scale Science on the United States, *Science* 134 (3473): 161-164.

•De Solla Price, D.J. (1963). *Little Science, Big Science.* New York, NY: Columbia University Press.

Note: "Little Science" is now referred to as "Small Science".



CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Data Storage Options

- Local
- Campus/Funder/Discipline-based
- Cloud-based

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Metadata

- Descriptive information
- Created for the purpose of retrieval & reuse
  - During creation of data
  - During archiving
- Standards
  - Data interoperability
  - Repository

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Metadata Standards

- Example Standards
  - Darwin Core (for biology data)
  - Data Documentation Initiative (DDI)
  - Directory Interchange Format
  - ISO 19115:2003
  - PREMIS
  - Science Data Literacy (SDL)
  - Seeing Standards (http://www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards_glossary_pamphlet.pdf)

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Metadata Quality Issues In Repositories

Examples Include:

- Missing critical fields
- Insufficient information
- Inconsistent field values & defined standards
- Fields with: Not available, n/a, none, & nbsp (space), wrong year, etc.
- Broken links
- Outdated contact information
- Challenges with updating
- etc.

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Digital Data Roles & Stakeholders

- Data Authors
- Data Scientists
- Data Managers
- Digital Curators
- Data Users
- Journal Publishers

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Different User Communities

- Observational users
- Infrastructure users
- Modelers
- Students
- Scientists
- Policy makers
- General public
- etc.

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Ethics, Legality & Copyright

- Data privacy, confidentiality, etc.
- Data disclosure
- Data ownership based on university & granting agency policies
- Data citation in development
  - MIT Libraries
  - Purdue's Online Writing Lab
  - APA Publication Manual, 6th edition
  - Dataverse Network Project

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Funder Requirements & Data Management Plans

- Data Plans Include:
  - The types of data to be produced
  - The standards that would be applied for format, metadata content, etc.
  - Provisions for archiving & preservation
  - Access policies & provisions
  - Plans for eventual transition or termination of the data collection after funding period
- Resources:
  - SHERPA – JULIET
  - UMN Funding Agencies & Data Management Guidelines
  - Digital Curation Centre Data Management Tool
  - NSF

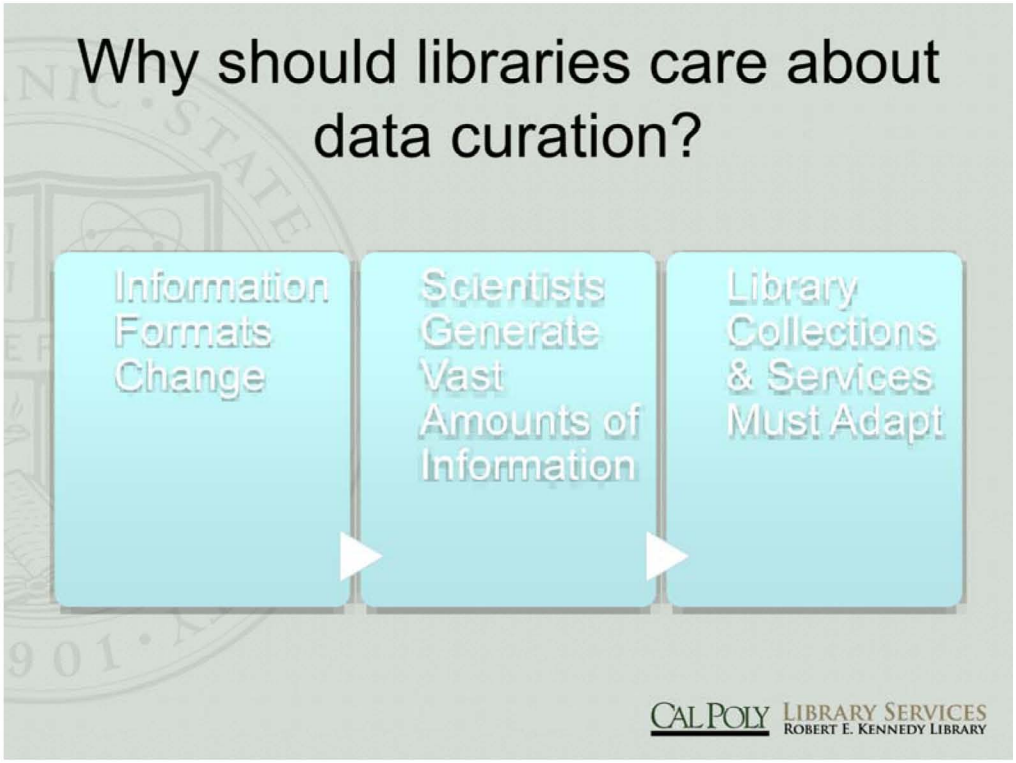
CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Data Plan Development

- What form and format is the data in?
- What is the expected lifespan of the dataset?
- How could the data be used, reused, and repurposed?
- How large is the dataset, and what is its rate of growth?
- Who are the potential audiences for the data?
- Who owns the data?
- Does the dataset include sensitive information?
- What publications or discoveries have resulted from the data?
- How should the data be made accessible?

From Purdue University Libraries, Conducting a Data Interview by Michael Witt and Jake Carlson

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY


Why should libraries care about data curation?
Information Formats Change
Scientists Generate Vast Amounts of Information
Library Collections & Services Must Adapt
CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

What knowledge do librarians bring to the table?

- LIS & archival theory
- Collection development, discovery, etc.
- LIS & IT partnerships
- Data management & scholarly communications
- Data organization

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Library Roles

- Self-education
- Outreach to scientists
- Provide new services –
  - Research,
  - Resources (repositories & databases)
  - Reference
- Collaborate with campus stakeholders

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Data Management Education

- Why Manage Your Data?
- Managing Data
- Back-up Practices
- Ethics, Legality & Copyright
- Funder Requirements & Data Plans
- Science Data & Repositories

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Managing Data Basics for Data Creators

1. Ensure data is accessible for the long-term, save a copy of data in a non-proprietary, commonly-accessible formats.
2. Keep a records of how data is produced & store it in a text file in the same directory as the data.
3. Use a folder/directory structure with a clear, documented naming scheme.
4. Use naming conventions for files
5. Lots of Copies Keeps Stuff Safe! Keep 3 copies of data in geographically distributed locations.

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Data & Libraries

Data & Libraries Overview 2006-2010

Gold, A. (2010) Data Curation and Libraries: Short-Term Developments, Long-Term Prospects Available. Earth and Space Science Informatics: Strategies for Improved Marine and Synergistic Data Access and Interoperability Session, 2009 American Geophysical Union (AGU) Fall Meeting, 14–18 December, San Francisco, CA.

Available online at:
http://digitalcommons.calpoly.edu/lib_dean/27

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Current Challenges in Curation

- Theory, policy, application, practice
- Relationships, data, practices, curation activities
- Emerging models, new divisions of labor and new roles
- Conceptualizing collections, observations, datasets, etc.
- Lifecycles ,selection and appraisal
- Continuity of access to usable and useful data, sustainable service models
- Resource allocation, limited infrastructure

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

Questions?

Jeanine Scaramozzino
jscaramo@calpoly.edu

Data Curation Management Guide
http://libguides.calpoly.edu/data

VALUE THE SCHOLAR IN EVERYONE

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY